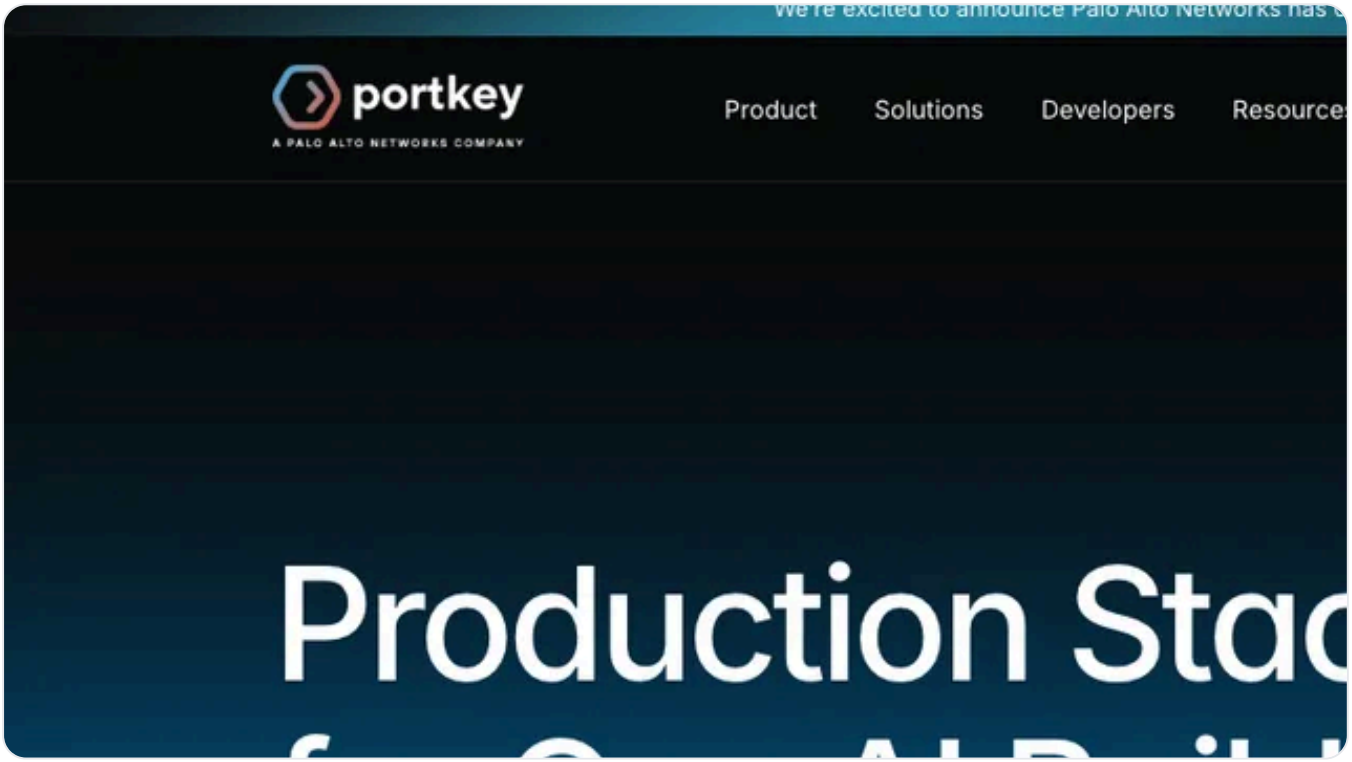
P

Portkey

AI gateway and production stack for GenAI teams, with observability, guardrails, governance, prompt management, and routing for large-scale LLM apps.



About

Portkey is an AI gateway for production-ready Gen AI apps that centralizes model access, routing, observability, guardrails, and governance. It is designed to help teams ship LLM-powered products across an organization, rather than wiring each app to models on its own. The platform also includes prompt management and usage monitoring, and the site describes it as open source with a free getting-started path and demo options for teams evaluating production deployment.

What is Portkey?

production without stitching together separate tools for model access, monitoring, and controls.

At a high level, Portkey is built around the idea of a central layer between your application and multiple LLMs. The company says it supports access to 1,600+ models, and the gateway is meant to simplify how teams route requests, add fallbacks, and manage usage across an organization. Because the platform also includes observability, guardrails, governance, and prompt management, it is not just a gateway in the narrow sense; it is an operational layer for LLM apps.

Portkey also states that Palo Alto Networks has completed its acquisition of the company. That is relevant context for buyers who care about vendor backing and enterprise direction.

Key features

Portkey's AI gateway is the core feature, but the platform bundles several production-oriented capabilities around it.

- **AI gateway**: One layer for connecting apps to many LLMs, rather than hard-coding provider logic into every service.
- **Model routing**: Requests can be directed across models, which is useful when you need control over latency, availability, or provider choice.
- **Fallbacks**: If a chosen model or path fails, the gateway can help teams design backup behavior.
- **Caching**: The source tagline highlights caching, which can reduce repeated work and support more efficient usage patterns.
- **Observability**: Portkey includes monitoring so teams can inspect model behavior and catch anomalies earlier.
- **Guardrails**: The platform mentions guardrails for safer Gen AI use, which suggests policy-based controls around requests or responses.
- **Governance**: Governance features are intended for org-wide control, especially when multiple teams use the same AI stack.
- **Prompt management**: Prompt tooling is included so teams can manage prompt changes centrally instead of scattering them across codebases.

Model catalog: The site also surfaces a model catalog, which helps users inspect available options from one place.

How to use it

Using Portkey usually starts with deciding whether you want a gateway for one app or a shared layer for several teams.

Review the docs or request a demo. The site offers both a docs path and a book-a-demo path, so you can choose self-serve or guided evaluation.

Connect your Gen AI app to the gateway. Portkey is meant to sit between your product and downstream LLMs.

Choose models and routing rules. Use the model catalog and routing layer to decide which LLM should handle each request.

Add caching and fallback behavior. These controls help with reliability and repeated traffic patterns.

Enable observability and guardrails. Track model behavior, monitor usage, and add policies before scaling to more users.

Manage prompts centrally. Keep prompt changes organized so teams can iterate without losing control.

That workflow makes Portkey feel less like a single API wrapper and more like an operating layer for a production AI gateway setup.

Who it is for

Portkey is best suited for teams that already build with LLMs and now need more control.

AI product teams that want a shared AI gateway instead of provider-specific code in every app.

Engineering orgs that need observability and governance across multiple Gen AI projects.

Platform teams looking for routing, caching, and fallback support in one layer.

Teams shipping to production that care about guardrails, usage monitoring, and prompt management.

If you only need a simple experiment or a one-off prototype, Portkey may be more capability than you need. Its value shows up when the AI gateway becomes part of a real production stack.

What to know before you use it

Portkey is powerful, but it is not an answer to every LLM challenge.

- It does **not** replace the need to choose suitable models, define prompts carefully, or test outputs in your own product.
- Observability and guardrails help, but they do **not** guarantee correct or safe model behavior.
- The platform supports production workflows, yet you still need to validate routing, fallback logic, and prompt changes in your own environment.
- Because the site mentions 1,600+ models, the exact models you need should be confirmed in the current catalog or docs rather than assumed.
- For official pricing, enterprise terms, or acquisition-related questions, the site's pricing and demo pages are better sources than third-party summaries.

In short, Portkey is an AI gateway for teams that want centralized control, production tooling, and a clearer path from prototype to deployment. If you are building Gen AI apps at scale, the AI gateway model is the main reason to evaluate Portkey.

FAQ

What is Portkey?

Is Portkey free to start with?

Is Portkey open source?

What does Portkey actually do for LLM apps?

Which platforms does Portkey support?

What makes Portkey suitable for production?

Who should use Portkey instead of a basic API wrapper?

How do I get started?

Does Portkey replace model testing and validation?

Sponsored

Belay – Awa... Featured
Developer Tools

Stays awake while your agents work.

Free # Developer Tools

View ↗

Partly Featured
Finance

Partly is a mobile app that lets friends quickly scan receipts, split any bill fairl...

Free # Finance

View ↗

Visit website ↗

portkey

Pricing	Freemium
Category	AI
Platform	Web, API
Launched	Aug 2026

P Portkey

Verified 1 months ago

TAGS

ai-gateway # model routing # prompt-management # guardrails # governance # llm # open source # gen-ai # caching # observability

© 2026 Dir Hub

P Portkey